Effects of the Achieve 3000 Program on the English Language Acquisition and English Proficiency Results of ELL Students

At a Glance

The 2014 CELLA and FCAT 2.0 reading outcomes were compared for students in the Program and Comparison Samples while adjusting for the students' 2013 results. It was found that the students in the Program Sample had higher 2014 CELLA and 2014 FCAT 2.0 reading adjusted mean scale scores than did students in the Comparison Sample for all but one comparison. However, the differences in the adjusted mean scale scores were small.

Because students in both student samples participated in the program during the 2013-2014 school year, the comparisons were made between students who differed on the *extent* of their participation in the program. Thus, the findings could be interpreted in the following way: the degree of program participation had a small positive effect on students' English language acquisition and reading achievement results.

Introduction

This evaluation was conducted at the request of the Division of Bilingual Education and World Languages, which requested that the Office of Assessment, Research, and Data Analysis examine the English language acquisition outcomes of students who participated in *Achieve 3000*, a computer based instructional program. *Achieve 3000* is a differentiated online instructional program intended to enable continued student reading growth. The program is designed to adjust reading materials to students' Lexile/reading levels, providing more challenging content as students' reading levels improve.

Evaluation Design

This section describes sampling procedures used to select schools and students. In addition, it addresses the outcome measures used in the evaluation and describes the data analyses performed.

Student Selection

Because *Achieve 3000* is implemented to in virtually all secondary schools and K-8 centers in the District, it was not possible to select a sufficiently large sample of students not participating in the program. Instead, all students in grades 6-12 who participated in the program during 2013-2014 were selected initially. Then, to allow for the sufficient exposure to the program activities, only students who participated in the program for at least 20 weeks were selected for achievement comparisons. This selection rule excluded about 15% of students with smaller lengths of participation. In addition, only the students with complete data on the 2013 and 2014 CELLA were used in the analysis.

Then, the length of student participation in the program was used to split the set of students into two approximately equal groups. The first group, for which the length of program participation during the 2013-2014 was at least 10.45 hours, constituted the Program Sample. The Comparison Sample included all students who participated in the program for less than 10.45 hours.

The program participation, demographic and English language proficiency characteristics of the two samples are shown in Table 1. The table shows that the student groups were reasonably well matched demographically. In terms of prior English language proficiency, students in the Program Sample had slightly higher mean scale scores in all three modalities on the 2013 Comprehensive English Learning Assessment (CELLA) for both grade-level groups. Although these differences were small compared to the standard deviations of scores, they were taken into account statistically as explained in the *Data Analysis* section.

Table 1

Program Participation, Demographic, and English Language Proficiency Characteristics of the two Samples

	Grades 6-8		Grades 9-12	
	Program (N=2879)	Comparison (N=3089)	Program (N=3164)	Comparison (N=2953)
Mean time of Program Participation	22.5	5.8	20.7	5.8
Mean numbers of Activities Completed	75.0	21.4	71.3	25.3
Percentage of Students who are				
Hispanic	84.9	83.0	80.2	83.1
Eligible for the federal free/reduced price lunch program	90.0	92.6	88.8	89.6
Mean Scale Scores on the 2013 CELLA (Standard Deviations)				
Listening/Speaking	716.8 (47.2)	715.7 (50.2)	722.5 (47.6)	718.1 (53.6)
Reading	727.5 (38.9)	723.9 (44.2)	746.9 (36.1)	740.7 (41.7)
Writing	714.6 (36.8)	711.7 (38.5)	721.1 (33.6)	715.9 (36.9)

Note: the figures in the table regarding the program participation and demographic characteristics of students in the two samples were practically identical across grade levels.

The 2013 reading achievement results of students in the two samples are shown in Table 2. The table shows that students in the Program Sample had slightly higher mean scale scores on the reading component of the 2013 FCAT 2.0 than student in the Comparison Sample for four out of five grade levels. In the data analysis, statistical adjustments for these differences were made.

Table 2

2013 Reading Achievement Results of the two Samples

Grade	Sample	2013 FCAT 2.0 Reading Scale Score	
		Mean	Standard Deviation
6	Program (N=1140)	197.2	16.4
	Comparison (N=1119)	197.6	16.7
7	Program (N=1002)	200.5	17.2
	Comparison (N=877)	197.5	17.6
8	Program (N=906)	202.9	16.0
	Comparison (N=832)	200.5	16.2
9	Program (N=816)	210.0	17.0
	Comparison (N=891)	205.9	16.5
10	Program (N=898)	213.3	17.6
	Comparison (N=913)	210.2	18.1

Outcome Measures

Student results on the 2013 and 2014 CELLA were used to examine the English language acquisition outcomes of students in the Program and Comparison Samples. CELLA is a four-skill language proficiency assessment that tests students' listening, speaking, reading, and writing skills. The results are provided as scale scores in the three domains: oral (listening/speaking), reading and writing. CELLA is administered at four different levels. For the analyses in this evaluation, 2014 CELLA results for levels C (grades 6-8) and D (grades 9-12) were used.

In addition, student results on the reading component of the FCAT 2.0 were compared for students in the Program and Comparison Samples. These analyses were carried out separately for each grade level.

Data Analysis

The General Linear Model (GLM) was used to compare the 2014 CELLA mean scale scores for students in the Program and Comparison Samples. The analyses were carried out separately for each of the grade level groupings of students, and independently for each of the three areas in which CELLA scale scores are reported: Listening/Speaking, Reading, and Writing. The 2013

CELLA scores in each modality were used as covariates while the program participation dichotomous indicator was used as a fixed factor.

In addition, the GLM analysis was performed to compare the students' mean scale scores on the reading component of the 2014 FCAT 2.0. In this analysis, the 2013 FCAT reading scores were used as a covariate. Prior to carrying out the GLM analysis, an assumption of the equality of slopes of regression lines of the 2014 CELLA outcomes on the 2013 outcomes had to be ascertained. It turned out that this assumption was violated in several cases of the CELLA outcome analyses. Because of that, an adjusted procedure comparing the 2014 CELLA outcomes for the low, medium, and high levels of the 2013 CELLA results was conducted. The low, medium, and high levels were defined as the mean minus one standard deviation, the mean, and the mean plus one standard deviation, respectively. All separate analyses were carried out at the .05 level of statistical significance.

Results

The results of the statistical analyses regarding the 2014 CELLA outcomes are presented in Table 3. In this table, the 2014 adjusted mean scale scores whose differences were found to not be statistically significant are shown in bold.

Table 3
2014 CELLA Adjusted Mean Scale Scores

The 2013 CELLA Achievement Level	Grades 6-8		Grades 9-12	
	Program (N=2879)	Comparison (N=3089)	Program (N=3164)	Comparison (N=2953)
	Listening/Speaking			
Low	716.8	711.1	709.8	703.7
Medium	743.2	738.7	742.0	738.0
High	769.5	766.4	769.5	767.2
	Reading			
Low	741.7	727.4	731.5	724.8
Medium	745.7	741.6	761.5	755.6
High	764.0	761.6	773.4	767.8
	Writing			
Low	719.2	712.7	710.5	706.3
Medium	734.9	730.2	733.4	729.5
High	**764.2**	**762.8**	766.3	762.8

Table 3 shows that the 2014 CELLA adjusted mean scale scores of students in the Program Sample were significantly higher than those of students in the Comparison Sample for all but one comparison. Specifically, the differences between the Program and Comparison Sample 2014 CELLA writing adjusted mean scale scores of students who scored at the high levels on the 2013 CELLA were not found to be statistically significant.

Table 4

2014 Reading Adjusted Mean Scale Scores

Grade	Sample	Mean Scale Score
6	Program (N=1140)	210.6
	Comparison (N=1119)	207.8
7	Program (N=1002)	211.2
	Comparison (N=877)	207.8
8	Program (N=906)	214.6
	Comparison (N=832)	211.2
9	Program (N=816)	219.2
	Comparison (N=891)	214.9
10	Program (N=898)	223.6
	Comparison (N=913)	221.3

Because the sample sizes were very large (at least 2879 in the case of CELLA and at least 816 in the case of the FCAT-based comparisons), even small differences are likely to be found statistically significant. Therefore, it is important to examine the differences that were found to be statistically different in terms of a standardized effect size, a measure of practical significance. In the analyses of CELLA data, the statistically significant differences expressed in terms of Cohen's d (the difference between the means expressed in standard deviation units) ranged from a low of .04 to a high of .12 in Listening/Speaking, from .06 to .17 in Reading, and from .10 to .17 in Writing. At the higher end of the range of effect sizes, the value of 0.17 can be interpreted as follows: a student in the Program Sample whose 2014 CELLA score corresponds to the adjusted mean for that group would have scored at the 57th percentile in the Comparison Sample assuming normality in the test score distribution. All these effect size estimates fall under .2, a value that is generally considered a small effect size.

In the analyses of the FCAT data, the values of standardized effect size varied from a low of .12 for students in grade 10 to a high of .21 for students in grade 9. These values are generally considered as indicators of a small effect size. A standardized effect size of .21 means that a student who scored at the mean in the Program Sample would have scored at the 58th percentile in the Comparison assuming the normal distribution of scores.

An investigation into the effects of the Achieve 3000 program was also conducted in May of 2013. The design of that study was virtually identical to that of the present one, and the findings of that evaluation were practically identical to those above. Specifically, the previous study also found statistically significant differences in performance between students in the Program and Comparison Samples on CELLA. However, the magnitude of those differences was small, ranging from .05 to .14 in terms of the standardized effect size.

Discussion

In this evaluation, the 2014 CELLA outcomes were compared for students in the Program and Comparison Samples while adjusting for the students' 2013 CELLA results. In addition, the 2014 FCAT 2.0 reading scores were compared. It was found that the students in the Program Sample had higher 2014 CELLA and 2014 FCAT 2.0 reading adjusted mean scale scores than did students in the Comparison Sample for all but one comparison. However, the magnitude of the effect size estimates was small.

It is important to note that students in both student samples participated in the program during the 2013-2014 school year. That is, the comparisons were made between students who differed on the extent of their participation in the program. Thus, the findings could be interpreted in the following way: the degree of program participation had a small positive effect on students' English language acquisition and reading achievement results.